DOCUMENT RESUME

ED 163 018 TM 007 313

AUTHOR Cook, Linda L.; Hambleton, Ronald K.
TITLE Application of Latent Trait Models to the Development of Norm-Referenced and Criterion-Referenced Tests.
PUB DATE [Mar 78]
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Complexity Level; Criterion Referenced Tests; *Item Analysis; *Mathematical Models; Norm Referenced Tests; *Test Construction; Test Items; Test Reliability
IDENTIFIERS Birnbaum 3 Parameter Logistic Model; Classical Test Theory; *Item Characteristic Curve Theory; Item Discrimination (Tests); *Latent Trait Theory; Test Theory

ABSTRACT

Latent trait models may offer considerable potential for the improvement of educational measurement practices, but until recently, they have received only limited attention from measurement specialists. This paper provides a brief introduction to latent trait models, and provides test practitioners with a non-technical introduction to the development of norm-referenced and criterion-referenced tests using latent trait models. Distinctions between standard test development procedures and test development procedures utilizing latent trait models are drawn. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED163018

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Linda L. Cook
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Application of Latent Trait Models to the Development of Norm-Referenced and Criterion-Referenced Tests

Linda L. Cook and Ronald K. Hambleton
University of Massachusetts, Amherst

Abstract

Latent trait models offer considerable potential for the improvement of educational measurement practices, but until recently, they have received only limited attention from measurement specialists. Reasons for the limited attention given to the models include their complexity and relative newness, and the shortage of technical guidelines for their use.

This paper has two purposes: (1) to provide a brief introduction to latent trait models, and (2) to provide test practitioners with a non-technical introduction to the development of norm-referenced and criterion-referenced tests using latent trait models. Toward accomplishing the second purpose, where relevant, distinctions between standard test development procedures and test development procedures utilizing latent trait models are drawn.

TM 007 313

Application of Latent Trait Models to the Development of Norm Referenced and Criterion-Referenced Tests[1,2,3]

Linda L. Cook and Ronald K. Hambleton
University of Massachusetts, Amherst

Standard testing and measurement technology has several shortcomings. For one, the values of standard item statistics are group-dependent. That is, typically they vary from one sample of examinees to another. Therefore, standard item statistics are only useful in test construction for examinee populations very similar to the sample of examinees in which the item statistics were obtained. A second shortcoming of standard testing technology is that comparisons of examinees on the ability measured by a set of test items comprising a test are limited to situations where examinees are administered the same (or parallel) test items. An additional limitation of standard testing technology is that it provides no basis for determining what a particular examinee might do when confronted with a test item. Such information is necessary, for example, if a test designer desires to predict test score characteristics in one or more populations of examinees or to design tests with particular characteristics for certain populations of examinees. For these and other reasons, many psychometricians have been investigating and developing other theories of mental measurements. Latent trait theory is one theory that is receiving considerable attention. (Sometimes it is referred to as item response theory, or item characteristic curve theory.)

---

[1]Paper presented at the annual meeting of NCME, Toronto, 1978.

[2]Laboratory of Psychometric and Evaluative Research Report No. 72. Amherst, MA: School of Education, University of Mass., Amherst, 1978.

[3]This project was performed pursuant to a contract from the United States Air Force Office of Scientific Research. However, the opinions expressed here do not necessarily reflect their position or policy, and no official endorsement by the Air Force should be inferred.

TM007 313

Latent trait models have been applied successfully to a variety of difficult measurement problems (for example, test score equating, item bias, and individualized testing). The practical value of latent trait models has been established clearly. Nevertheless, specific guidelines for applying latent trait models to particular measurement problems are still in the developmental stages and therefore the practitioner's task is not as easy as would be desirable.

The goal of this paper is to introduce test practitioners to the topic of test development utilizing latent trait models. Specifically, the paper was prepared to achieve two purposes:

1. To provide a brief introduction to latent trait models, and
2. to provide test practitioners with a non-technical introduction to the development of norm-referenced and criterion-referenced tests using latent trait models.

Toward accomplishing the second purpose, where relevant, distinctions between standard test development procedures and test development procedures utilizing latent trait models are drawn.

This paper is the first in what we expect will be a series of papers prepared for the Air Force Office of Scientific Research. The thrust of the papers will be to provide detailed guidelines for the selection of latent trait models for specific intended uses of test scores, for estimating item and ability parameters, and for test design and analysis. Most discussions of the test development process, have involved the one-parameter logistic model [see for example, Hambleton, Swaminathan, Cook, Eignor, and Gifford, 1978; Rentz and Rentz, 1978]. We do not intend to debate the effectiveness of one latent trait model over another for

developing tests in this paper, but it does seem to us that the three-parameter logistic model holds considerable potential. Therefore, a general overview of latent trait theory as it applies to test development, using the three-parameter logistic model as a frame of reference, is presented.

Introduction to Latent Trait Theory

A theory of latent traits supposes that, in testing situations, examinee performance on a test can be predicted (or explained) by examinee characteristics, referred to as traits. Scores for examinees on these traits are estimated and used to predict or explain test performance (Lord and Novick, 1968). Since the traits are not directly measurable, they are referred to as latent traits or abilities. A latent trait model specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test. The relationship between the "observable" and the "unobservable" quantities is described by a mathematical function.

When selecting a particular latent trait model to apply to one's test data, it is necessary to consider whether the data satisfy the assumptions of the model. If they do not, different test models should be considered. Alternately, some psychometricians (for example, Wright, 1968) have recommended that test developers design their tests so as to satisfy the assumptions of the particular latent trait model they are interested in using. In this way, the advantages of the particular latent trait model of interest can be utilized.

The three fundamental assumptions underlying the most commonly used

latent trait models are: The unidimensionality of the test items, local independence, and the mathematical form of the item characteristic curves. Each of these assumptions will be discussed briefly. Two other important topics will also be considered: Item and test information curves, and efficiency.

The Assumption of Unidimensionality

The assumption of a unidimensional set of test items is a common one for test constructors, since they usually desire to construct unidimensional tests so as to enhance the interpretability of a set of test scores (Lumsden, 1976).

Lumsden (1961) provided an excellent review of methods for constructing unidimensional tests. He concluded that the method of factor analysis held the most promise. Fifteen years later he reaffirmed his conviction (Lumsden, 1976). Essentially, Lumsden recommends that a test constructor generate an initial pool of test items selected on the basis of empirical evidence and a priori grounds. Such an item selection procedure will increase the likelihood that a unidimensional set of test items within the pool of items can be found. If test items are not preselected, the pool may be too heterogeneous for the unidimensional set of items in the item pool to emerge. In Lumsden's method, a factor analysis is performed and items that are not measuring the dominant factor obtained in the factor solution are removed. The remaining items are factor analyzed, and again, "deviant" items are removed. The process is repeated until a satisfactory solution is obtained. Convergence is most likely when the initial item pool is carefully selected to include only items that appear to be measuring a common trait. Lumsden proposed that

the ratio of first factor variance to second factor variance be used as an "index of unidimensionality."

Local Independence

The second assumption is that of local independence. The assumption states that the test item responses of a given examinee are statistically independent. This means that an examinee's performance on one item does not affect his or her performance on other items in the test. The result would be obtained if the test items measure a single ability.

Item Characteristic Curves

An item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the set of items contained in the test. There is no concept comparable to the notion of an item characteristic curve in standard test technology. A primary distinction among different latent trait models is in the mathematical form of the corresponding item characteristic curves. It is up to the user to choose one of the many mathematical forms for the shape of the item characteristic curves. In doing so, an assumption about the items is being made which can be verified later by how well the chosen model "explains" obtained test results.

Each item characteristic curve for a particular latent trait model is a member of a family of curves of the same general form. The number of parameters required to describe an item characteristic curve will depend on the particular latent trait model. We will restrict our discussion in this paper only to those item characteristic curves that are described by a three parameter logistic model.

The mathematical expression for the three-parameter logistic curve is:

$$P_g(\theta) = c_g + (1-c_g)\frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}}, \quad g=1, 2, \ldots, n,$$

where:

$P_g(\theta)$ = the probability that an examinee with ability level θ answers item g correctly,

b_g = the item difficulty parameter,

a_g = the item discrimination parameter,

and

D = 1.7 (a scaling factor).

The parameter c_g is the lower asymptote of the item characteristic curve and represents the probability of examinees with low ability correctly answering an item. The parameter c_g is included in the model to account for test response data at the low end of the ability continuum, where among other things, guessing is a factor in test performance. It is now common to refer to the parameter c_g as the pseudo-chance level parameter in the model.

Typically, c_g, assumes values that are smaller than the value that would result if examinees of low ability were to guess randomly to the item. As Lord (1974a) has noted, this phenomenon can probably be attributed to the ingenuity of item writers in developing "attractive" but incorrect choices. For this reason, avoidance of the label "guessing parameter" to describe the parameter c_g is desirable.

Figure A shows S shaped curves representing two, three-parameter logistic curves. These curves differ in slope, translation and lower asymptote. The differences in slope are due to differences in the "a_g"

parameters; differences in translation are due to differences in the "b_g" parameters; and differences in the "c_g" parameters account for the differences in the lower asymptotes of the two curves.

Figure 1. Two typical three-parameter logistic curves.

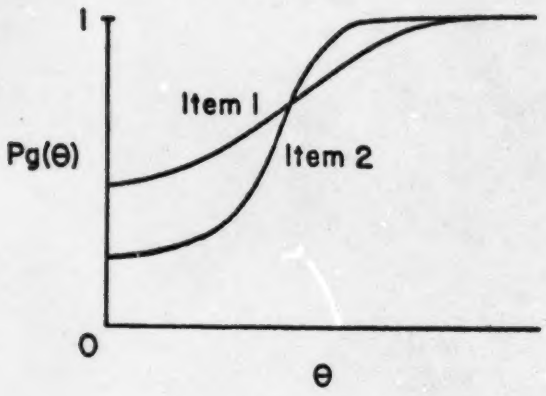

Table 1 presents the parameters for 18 test items, and probabilities of success associated with seven ability levels, for each of the test items. A study of the table will provide the practitioner with some insights into the effects of varying the three-parameters, a_g, b_g, and c_g. In a few words, high values of a_g result in "steeper" curves than low values. Other parameters being equal, examinees at any ability level will do better on items with lower b_g values. Finally, other parameters being equal, examinees will have higher probabilities of success on items where $c_g > 0$.

Table 1

Item Characteristic Curves for
a Set of Typical Test Items

Item Parameters				Probabilities[1]						
				Ability Level						
Number (g)	Difficulty (b_g)	Discrimination (a_g)	Pseudo Chance (c_g)	-3	-2	-1	0	1	2	3
1	-1.5	.19	.00	38	46	54	62	69	76	81
2	0.0	.19	.00	28	34	42	50	58	66	72
3	1.5	,19	.00	19	24	31	38	46	54	62
4	-1.5	.59	.00	18	38	62	82	92	97	99
5	0.0	.59	.00	05	12	27	50	73	88	95
6	1.5	.59	.00	01	03	08	18	38	62	82
7	-1.5	.99	.00	07	30	70	93	99	100	100
8	0.0	.99	.00	01	03	16	50	84	97	99
9	1.5	.99	.00	00	00	01	07	30	70	93
10	-1.5	.19	.25	54	59	66	71	77	82	86
11	0.0	.19	.25	46	51	56	63	69	74	79
12	1.5	.19	.25	39	43	48	54	59	66	71
13	-1.5	.59	.25	39	53	72	86	94	98	99
14	0.0	.59	.25	29	34	45	63	80	91	96
15	1.5	.59	.25	26	27	31	39	53	72	86
16	-1.5	.99	.25	31	48	77	94	99	100	100
17	0.0	.99	.25	25	28	37	63	88	97	100
18	1.5	.99	.25	25	25	26	31	48	77	94

[1]Decimal points have been omitted.

Item and Test Information Curves

Once a latent-trait model is specified, the precision with which it estimates examinee ability can be determined. Birnbaum (1968) defined the notion of information as a quantity inversely proportional to the squared length of the confidence interval around an estimate of an examinee's ability. The standard error of estimate of ability is equal to $1/\sqrt{\text{information}}$. When information at an ability level is high, we have narrow confidence bands around our estimates. If information is low, we have wider confidence bands. Because the information function varies with ability level, it has been suggested that test information curves ought to replace the use of classical reliability estimates and standard errors of measurement in test score interpretations.

In mathematical terms, Birnbaum (1968) gives the information curve of a given scoring formula by

$$I_y(\theta) = \frac{\left(\sum_{g=1}^{n} w_g P'_g\right)^2}{\sum_{g=1}^{n} w_g^2 P_g Q_g}.$$

In the expression above, $I_y(\theta)$ is the amount of information at ability level θ provided by the scoring formula y, where

$$y = \sum_{g=1}^{n} w_g X_g;$$

the variable X_g is 0 or 1 depending on whether or not item g is answered correctly; P_g is the probability of a correct answer to item g by an examinee with ability level θ; Q_g is equal to $1 - P_g$; P'_g is the slope of the item characteristic curve at ability level θ; and the item scoring weights are w_g, $g = 1, 2, \ldots, n$.

Birnbaum (1968) has shown that the maximum value of $I_y(\theta)$, referred to as the test information curve, is given by

$$I(\theta) = \sum_{g=1}^{n} \left(\frac{P'^2_g}{P_g Q_g}\right) . \qquad [1]$$

The maximum value of the information curve of a given scoring formula is obtained when the scoring weights are chosen, such that

$$w_g = \frac{P'_g}{P_g Q_g} .$$

The quantity P'^2_g/P_gQ_g is the contribution of item g to the information function of the test and is referred to as the item information function. Item information functions have an important role in determining the accuracy with which ability is estimated at different levels of θ. Each item information curve depends on the slope of the particular item characteristic curve and the conditional variance of test scores at each ability level θ. The higher the slope of the item characteristic curve and the smaller the conditional variance, the higher will be the item information curve at that particular ability level. The height of the item information curve at a particular ability level is a direct measure of the usefulness of the item for precisely measuring ability at that level.

Table 2 lists the amount of information, as it varies with ability level, for each of the 18 items found in Table 1. It is easy to see that items with low discrimination parameters, such as items 1-3 and 10-12, provide very little information at all ability levels. As is to be expected, difficult items provide maximum information at high ability levels. Of course the opposite is true for easy items. The effect of guessing on the amount of information is apparent from a comparison of items 7-9 with

Table 2

Item Information Curves for a Set of Typical Test Items

Item Parameters				Item Information[1]						
				Ability Level						
Number (g)	Difficulty (b_g)	Discrimination (a_g)	Pseudo Chance (c_g)	-3	-2	-1	0	1	2	3
1	-1.5	.19	.00	02	03	03	02	02	02	02
2	0.0	.19	.00	02	02	03	03	03	02	02
3	1.5	.19	.00	02	02	02	02	03	03	02
4	-1.5	.59	.00	15	24	24	15	07	03	01
5	0.0	.59	.00	15	11	20	25	20	11	05
6	1.5	.59	.00	01	03	07	15	24	24	15
7	-1.5	.99	.00	19	60	60	19	04	01	00
8	0.0	.99	.00	02	09	37	71	37	09	02
9	1.5	.99	.00	00	01	04	19	60	60	19
10	-1.5	.19	.25	01	02	02	02	02	01	01
11	0.0	.19	.25	01	01	01	02	02	02	01
12	1.5	.19	.25	01	01	01	01	02	02	02
13	-1.5	.59	.25	05	13	15	11	05	02	01
14	0.0	.59	.25	01	03	09	15	14	08	03
15	1.5	.59	.25	00	00	01	05	13	15	11
16	-1.5	.99	.25	04	28	40	14	03	01	00
17	0.0	.99	.25	00	01	12	42	27	07	01
18	1.5	.99	.25	00	00	00	04	28	40	14

[1]Decimal points have been omitted.

items 16-18. These two sets of items are identical, with the exception of the value of the guessing parameters. It is obvious that guessing dramatically reduces the amount of information obtained from a particular item at nearly all ability levels.

The information function for the test composed of the items is obtained by summing the ordinates of the item information curves. From Equation [1] it is clear that items contribute independently to the test information function. Birnbaum (1968) has also shown that with his three-parameter model, an item provides maximum information at an ability level θ, where

$$\theta = b_g + \frac{1}{1.7\, a_g} \log_c 1/2\ (1 + \sqrt{1 + 8c_g}).$$

If guessing is minimal, then $c_g = 0$, and $\theta = b_g$. When $c_g > 0$, the point of maximum information is shifted to the right of the item difficulty value, b_g.

Efficiency

A concept closely related to test information is the concept of efficiency. An efficiency curve is formed by calculating the ratio of two information curves at different points on an ability continuum. The efficiency curve provides a measure of the relative effectiveness of two tests (each characterized by a different information curve) for measuring ability. In test development work it is common to compare the efficiency of different test designs (i.e., tests composed of different items) for measuring ability at different locations on the ability continuum. Whereas the shapes of test information curves depend on the metric chosen for measuring ability, efficiency curves do not, and therefore they are particularly useful in test development work.

The process of determining the relative efficiency of two tests is employed more often as part of the analysis of existing tests than as a part of the test development process (see for example, Lord, 1974b). The distinction between test analysis and test development has been made by Rentz and Bashaw (1977). Basically they define the test development process as one that allows items that do not fit the model to be discarded, whereas in test analysis applications, "the model becomes fixed and data are in effect 'fitted' to it." The distinction made by Rentz and Bashaw between test development and test analysis is a useful one.

Development of Tests Utilizing Latent Trait Models

The test development process consists of the following steps:

1. establishing test specifications,
2. preparing the item pool,
3. determining item characteristics,
4. selection of items,
5. norming (for norm-referenced tests),
6. specification of cutting scores (for criterion-referenced tests),
7. reliability studies,
8. validity studies, and
9. final test compilation.

Regardless of the test development model employed, the nine steps above must be addressed. The important differences between developing tests using standard methods and latent trait models occur at steps 3, 4, and 7. The following discussion will center on these three steps.

Determining Item Characteristics

Standard item analysis techniques involves three basic steps: (1) determining item difficulty; (2) determining item discrimination, and (3) the detection of bad items. The major problem with the standard approach is that item statistics are not sample invariant. This issue is well-known and need not be discussed further here. One advantage of the standard approach is that estimation of item parameters (difficulty and discrimination) is straightforward and requires only a moderate sample size for obtaining stable parameter estimates.

Detection of bad items (for norm-referenced tests at least) is basically a matter of studying item statistics. Items that are too easy or too difficult or non-discriminating are usually deleted.

The process is quite different when latent trait models are employed to execute the item analysis. As mentioned previously, the major advantages of latent trait methods is that they lead, in theory, to item parameters that are sample invariant. Difficulties that have been cited are the necessity for large sample sizes in order to obtain stable item parameter estimates and the mathematical complexity of the techniques used to obtain these estimates.

The detection of bad items is not as straightforward as when standard techniques are employed. Items are generally evaluated in terms of their goodness of fit to the model using a chi-square statistic. Rentz and Rentz (1978) and Rentz and Ridenour (1968) have given an excellent reviews of research related to goodness of fit measures used with the one-parameter logistic model. Many of their observations and interpretations of the research are applicable to the use of these measures with the three-parameter logistic model as well.

Unfortunately, there are problems involved in applying the chi-square statistic. The basic problem concerns the fact that the chi-square statistic is extremely sensitive to sample size, i.e., if enough observations are taken, the null hypothesis that the model fits the data will always be rejected using the chi-square test. Rentz and Rentz (1968) suggest caution on the part of the practitioner when interpreting the results of a goodness of fit evaluation. They suggest that relative sizes be considered rather than absolute values. They also suggest that the practitioner disregard any probability values associated with the test statistic due to the previously mentioned fact that if sample size is sufficiently large, probabilities will almost always be less than the critical value set by the investigator. These recommendations seem eminently sensible.

In summary, the item analysis process, when employing standard test development techniques, consists of the following steps: (1) determining sample-specific item parameters employing simple mathematical techniques and moderate sample sizes; and (2) deleting items based on statistical criteria. In contrast, item analysis using latent trait models involves: (1) determining sample-invariant item parameters using relatively complex mathematical techniques and large sample sizes; and (2) utilizing goodness of fit criteria to detect items that do not fit the specified latent trait model.

Item Selection

When applying standard test development techniques to the construction of norm-referenced tests, items are selected on the basis of their two characterizing parameters, item difficulty and item discrimination. An attempt is always made to choose items with the highest

discrimination parameters. The choice of level of item difficulty is usually governed by the purpose of the test and the anticipated ability distribution of the group the test is intended for.

Most norm-referenced achievement tests are commonly designed to differentiate examinees with regard to their competence in the areas tested, i.e., the test is designed to yield a broad range of scores maximizing discrimination among all examinees taking the test. When a test is designed for this purpose, items are generally chosen to have a medium level and narrow range of difficulty.

It may be that the purpose of a test is to select a small group of high ability examinees. A scholarship examination is a good example. In this situation, items are generally selected such that an examinee whose ability places him/her at exactly the cutting point would have a probability of .5 of answering that item correctly.

The important point to note is that because standard item parameters are not sample invariant, the success of the technique described above depends directly on how closely the sample used to determine the item parameters, employed in the item selection process, matches the population for which the test is intended.

Latent trait theory affords the test developer a far more powerful method of item selection. Of course, item selection is, as when standard methods are employed, based on the intended purpose of the test. However, items are selected depending upon amount of information they contribute to the total amount of information supplied by the test.

Lord (1971) has outlined a procedure, originally conceptualized by Birnbaum (1968) for the use of item information curves in the test

building process. Basically, this procedure entails the following steps:

1. Describe the shape of the desired test information curve. Lord (1977) calls this the target information curve.
2. Select items with item information curves that will fill up the hard-to-fill areas under the target information curve.
3. After each item is added to the test, calculate the test information curve for the selected test items.
4. Continue selecting test items until the test information curve approximates the target information curve to a satisfactory degree.

It is obvious that the use of item information curves in the manner described above will allow the test developer to produce a test that will very precisely fulfill any set of desired test specifications.

The following two examples are intended to demonstrate how the practitioner would apply Birnbaum's technique to develop, first, a typical norm-referenced achievement test (one designed to maximize the range of examinee scores) and secondly, a scholarship exam (designed to yield maximum discrimination among high ability examinees).

The items used to illustrate Birnbaum's procedure are those found in Tables 1 and 2. Although somewhat typical, this set of items is far from being representative of an ideal item pool. The discrimination parameters of all items are fairly low, therefore most of the items afford only a limited amount of information. Those items with discrimination parameters of .19 supply no usable information at all and so can be eliminated from our consideration. Elimination of the six items with discrimination parameters of .19 leaves an item pool of 12. Of course, no practitioner would attempt to develop a test with a pool of only 12 items, but, the number of items is sufficient for illustrative purposes.

The first example we will discuss is the development of a typical norm-referenced achievement test. The initial consideration is Birnbaum's first step, the establishment of the target information curve.

The practitioner would most probably desire a test of this type to provide maximum discrimination or information in the ability range of -2 to +2, i.e., as abilities are scaled to have a mean of zero and a standard deviation of one, this ability range would contain approximately 95% of examinee abilities (if ability is normally distributed). Once the area of maximum test information is determined, the next decision to be made is the accuracy with which it is desired to estimate abilities within this range. Suppose the test developer could tolerate an error of .33 (standard errorof estimation = .33).

Since the standard error of estimation (SEE) equals $1/\sqrt{\text{Information}}$, if SEE = .33, then Information $\doteq$ 9. Thus, to obtain estimates of ability to the desired degree of precision across the ability scale (from -2.0 to +2.0) items must be selected from the item pool to produce a test information curve (height $\doteq$ 9) from -2.0 to +2.0 on the ability scale. The tails of the target information curve, those sections below -2 and above +2, are of no interest to the test developer and can take on any form.

After the target information curve is specified, the next step is to select items with ordinates that, when summed, approximate the target information curve as closely as possible. Trial and error is involved in this step of the process. It is important that the test developer select the fewest possible items to fulfill the test requirements. For this particular example, the most useful items would be those supplying

a reasonable [illegible] over a wide ability range, e.g., items 5, [illegible] 7 or 9, which supply relatively more information [illegible] range of abilities, will be reserved to fill "hard to fill" areas under the target information curve.

Suppose the test developer began by selecting 10 items identical to item eight, the amount of information supplied at each ability level would be the following:

Ability Level	Test Information Curve (10 items)
-3	.20
-2	.90
-1	3.70
0	7.10
1	3.70
2	.90
3	.20

It is obvious that additional information is required at ability levels -2, -1, 1 and 2. Appropriate choices of items to fill the problem areas under the target information curve would be items similar statistically to 7 and 9. Suppose six items identical to item 7, and 6 identical to item 9 were chosen. The test information curve would now supply the following amount of information at the various ability levels:

Ability Level	Test Information Curve (22 items)
-3	1.34
-2	4.56
-1	7.54
0	9.38
1	7.54
2	4.56
3	.32

At this point, areas under the target information curve that are still deficient are those corresponding to ability levels -2, -1, 1 and 2. What is required to fill these areas are several easy and difficult items with fairly high discrimination parameters (items which are not contained in the item pool). To continue to attempt to fill these areas with the existing items would probably result in the final test being much longer than would be necessary if appropriate items were available.

The target information curve for a test designed as a scholarship examination would be one that produced information at the high end of the ability scale. Again, if the test designer was satisfied with estimating ability to an accuracy of $\pm$ .33, the information level desired in this area would be approximately equal to nine. The target information curve for this type of test would be one with an ordinate of nine at high levels of ability. The height of the target information curve at other points on the ability scale is not of interest. Items would be selected to fulfill the requirements of this test in the manner previously described.

Test Reliability

When standard test development methods are employed, one or more of the following approaches to reliability are used: (1) parallel-form reliability; (2) test-retest reliability; and (3) corrected split-half reliability. All three measures of reliability are sample specific. This unfortunate property of standard estimates of reliability reduces their usefulness. Another problem is that the well-known classical model estimates of reliability lead to a group estimate of error in individual

test scores. It is called the standard error of measurement. Intuitively, it does not seem to be reasonable to assume that the size of errors in examinee test scores is unrelated to the "true scores" of examinees taking the test.

The latent trait theory analog of test score reliability and the standard error of measurement is the test information function. The concept of information was described earlier in the paper and its use in the selection of test items was illustrated for the two examples given in the previous section. The use of the test information curve as a measure of accuracy of estimation is appealing for at least two reasons: (1) Its' shape depends only on the items included in the test, and (2) it provides an estimate of the error of measurement at each ability level.

An excellent discussion of the use of test information curves as a replacement for the traditional concepts of reliability and the standard error of measurement is given by Samejima (1977).

Some Final Comments

The viability of latent trait models for test development work is clear but more effective implementation could be achieved if several questions were satisfactorily answered.

The choice of a model is one question. At the test development stage, the practitioner has the option of developing items to fit a specific latent trait model. It would greatly facilitate the test development process, if practical guidelines existed that provided a logical basis for making this choice.

A second question concerns the reason for item misfit. At the present level of technical sophistication, the test developer, faced

with a misfitting item, can do little more than subjectively examine the item and hope that the reason for misfit will be apparent.

The problem of determining whether or not a pool of items can be considered unidimensional in an important one. Factor analytical techniques are often used for this purpose but there are problems (Hambleton et al., 1978; Lord and Novick, 1968).

Numerous test developers are now considering the use of latent trait models in their work. Hopefully this paper will provide some newcomers to the area with a suitable introduction to the topic.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: A review of models, technical issues, and applications. Review of Educational Research, in press.

Lord, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology, Vol. II. San Francisco: Freeman, 1974. (a)

Lord, F. M. Quick estimates of the relative efficiency of two tests as a function of ability level. Journal of Educational Measurement, 1974, 11, 247-254. (b)

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Lumsden, J. The construction of unidimension tests. Psychological Bulletin, 1961, 58, 122-131.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

Rentz, R. R., & Bashaw, W. L. The National Reference Scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.

Rentz, R. R., & Rentz, C. C. Does the Rasch model really work? A synthesis of the literature for practitioners. Mimeo, 1978.

Rentz, R. R., & Ridenour, S. E. The fit of the Rasch model to achievement tests. A paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg, VA, March 1978.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-248.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1968.